32nd Annual International Conference of the IEEE EMBS
Buenos Aires, Argentina, August 31 - September 4, 2010

Graph theory-based measures as predictors of gene morbidity

Raimon Massanet-Vila[1,2,3], Pere Caminal[1,2,3], Alexandre Perera[1,2,3]

[1]Dept. ESAII, Technical University of Catalonia (UPC), Barcelona, Spain.
Email: {raimon.massanet, pere.caminal, alexandre.perera}@upc.edu

[2]Biomedical Engineering Research Center (CREB), Barcelona, Spain.

[3]CIBER-BBN in Bioengineering, Biomaterials and Nanomedicine, Spain.

Abstract— **Previous studies have suggested that some graph properties of protein interaction networks might be related with gene morbidity. In particular, it has been suggested that when a polymorphism affects a gene, it is more likely to produce a disease if the node degree in the interaction network is higher than for other genes. However, these results do not take into account the possible bias introduced by the variance in the amount of information available for different genes. This work models the relationship between the morbidity associated with a gene and the degrees of the nodes in the protein interaction network controlling the amount of information available in the literature. A set of 7461 genes and 3665 disease identifiers reported in the Online Mendelian Inheritance in Man (OMIM) was mined jointly with 9630 nodes and 38756 interactions of the Human Proteome Resource Database (HPRD). The information available from a gene was measured through PubMed mining. Results suggest that the correlation between the degree of a node in the protein interaction network and its morbidity is largely contributed by the information available from the gene. Even though the results suggest a positive correlation between the degree of a node and its morbidity while controlling the information factor, we believe this correlation has to be taken with caution for it can be affected by other factors not taken into account in this study.**

I. INTRODUCTION

High throughput protein interaction identification methods, like yeast two-hybrid [1], high-throughput mass-spectrometry protein complex identification (HMS-PCI) [2], tandem affinity purification (TAP) [3], correlated mRNA expression and others, have allowed in recent years the building of large protein-protein interaction (PPI) networks, of a relatively high reliability. Although graphs have some limitations when modeling PPI networks, they have been widely used to model these interaction networks [4], [5]. In addition, graph theory has been applied to study PPI networks helping to unveil some of their characteristic network properties. Particularly, a great effort has been directed towards the discovery of relationships between graph properties of PPI networks and the morbidity of genes.

One of the most studied graph properties of PPI networks is the node degree. The degree of a node in a graph is the number of adjacent nodes. In a PPI context, the degree of a protein is the number of other proteins with which it interacts. Some authors have suggested that morbidity is related to node degree in interaction graphs [6], [7]. The idea behind this statement is that mutations in highly connected nodes could cause a major disruption of the network. Jeong et al. stated that PPI networks, like other real-world networks, have a scale-free topology, with few high-degree nodes and many low-degree nodes, with the degree distribution following a power law [8]. This kind of networks are known to increase the robustness of the network to random errors. However, these networks are vulnerable to errors in hub nodes (nodes of high degree). Other studies have suggested that the degree of nodes in PPI networks could be associated with the lethality of genes, with lethal gene mutations having higher degrees that non-lethal genes mutations in their graph representation [6]. Furthermore, it has been suggested that lethal genes could correspond to high-degree nodes that also disconnect the PPI network upon removal [7]. These results support the idea of morbidity of genes being a consequence of their central role in the proteomic network, independently of their biological function.

On the other hand, genes of known morbidity, along with their neighborhood, tend to be more thoroughly studied, in order to find additional modulating or cross-effect genes. This could cause a bias in the amount of PPI information available for genes, with morbid genes having more interactions reported than non-morbid genes, just because of the attention that the scientific community has payed to them. This could contribute in a causal effect between gene morbidity and node degree, and not the other way.

This work intends to further explore the relationship between gene morbidity and node degree, taking into account the number of publications for the different genes. In order to properly study this relationship, variance of node degrees should be explained by controlling the variance corresponding to the information that has been published about the genes they represent.

In this contribution, this is approached through a linear model that relates gene morbidity and node degree, detaching the variance caused by the varying amount of information.

This work was supported by the Spanish Ministerio de Educación y Ciencia under the Ramón y Cajal Program and TEC2007-63637/TCM and by the ISCIII under the CIBER initiative.

This amount of information is defined as the number of publications for each gene.

II. MATERIALS AND METHODS

The Online Mendelian Inheritance in Man (OMIM) database was used to obtain an estimation of the morbidity of a gene [9]. OMIM's data maps every reported human disease to the set of gene symbols that have been discovered to cause or modulate the anomaly. The morbid map used in this work was retrieved from OMIM in Feb-05 2010. This data maps 7461 different genes to 3665 OMIM identifiers.

The Human Proteome Resource Database (HPRD) was used to obtain protein interaction information [10]. This data was retrieved from the HPRD website, version of Jul-06 2009, and transformed into an undirected graph of 9630 nodes and 38756 interactions.

A set of software tools were written to automatically query PubMed web service in order to obtain an estimation of the amount of information avaialable for a given gene. This amount of information was estimated as the number of different publication identifiers obtained when querying single genes.

Of the 9630 nodes of the HPRD graph, 9374 could be mapped to a gene symbol. For each mapped symbol the following three measures were calculated: the node degree in the protein interaction graph, the number of OMIM identifiers (morbidity) and the number of PubMed identifiers associated (amount of information). In order to study the relationship between morbidity and degree, two samples were compared. The case sample consisted of the degrees of the 1873 genes that had at least one OMIM identifier associated (morbid). The control sample consisted of the degrees of equally sized randomly generated samples of non-morbid genes. The difference between the two populations was measured through a Mann-Whitney test, the null hypothesis being that the degrees of the two samples were equally distributed [11].

To further study the influence on the morbidity of a gene, the average degree and amount of information were calculated for each morbidity value. Then a linear model was built in order to quantify this influence. The morbidity was used as response variable, while amount of information and degree were used as explicative variables. This model is described by the following equation:

$$M = \alpha_1 \bar{I}_M + \alpha_2 \bar{D}_M + \beta \tag{1}$$

where M are the different values of morbidity, $\bar{I}_M$ is the average number of publications for each value of M and $\bar{D}_M$ is the average degree for each value of M.

All the mining and computing steps were performed using the R statistical programming language [12].

III. RESULTS

The results show statistically significant differences between degrees of morbid genes and degrees of non-morbid genes (see Fig. 1), with a maximum p-value of $6.72e-10$. This low value indicates that the null hypothesis does not hold, and therefore the degrees of morbid genes are higher than the degrees of non-morbid genes. This suggests that the more interactions a gene has the more likely it is to be related with diseases, which is coherent with previous results [6]. Even though this result seems intuitive, the effect of the varying amount of information available on the different genes is not negligible.

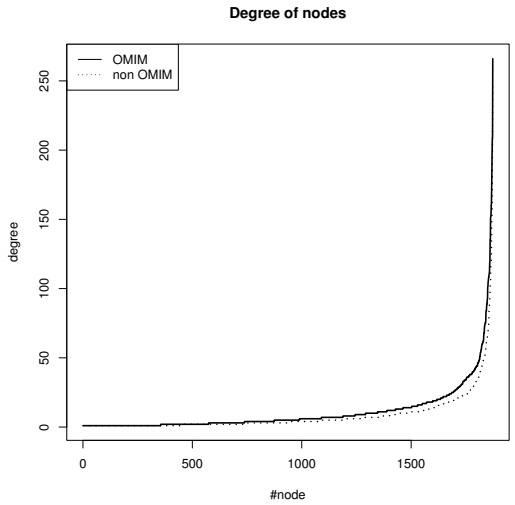


Fig. 1: Node degree distribution of genes in OMIM (continuous line) and genes not in OMIM (dashed line). The differences where found to be statistically significant, with a maximum p-value of $6.72e^{-10}$.

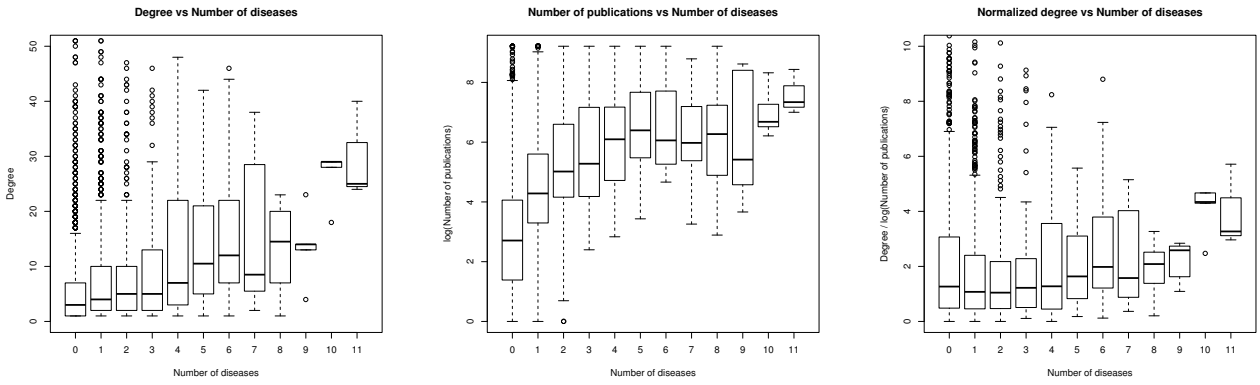
Genes were grouped according to the number of diseases they are related with. Fig. 2a shows the distribution of the node degrees with respect to the gene morbidity. Despite a low correlation value of 0.20, a positive correlation seems evident. However, this correlation could be affected by the fact that genes related to diseases tend to be more studied. Fig. 2b shows the distribution of the number of publications for the genes with respect to their morbidity. A very strong positive correlation seems obvious in this case too. In fact, the correlation value 0.26 is higher than the correlation with the degree.

In order to assess the effect of the variance in the amount of information available for different genes, we normalized the degree of the nodes, dividing it by the amount of publications in which they appear:

$$D_n(g) = \frac{D(g)}{I(g)} \tag{2}$$

where $D(g)$ is the degree of gene g and $I(g)$ is the amount of information (number of publications) available for gene g. Fig. 2c shows that when the number of interactions is normalized in such a way the positive correlation with the number of diseases is not as evident and the correlation value drops to -0.12.

The results for the model built for Equation 1 are shown in Tables I and II. The normally-distributed residuals (p-value



(a) Average degree of nodes as a function of the number of diseases related to them.

(b) Average number of publications (in logarithmic scale) of nodes as a function of the diseases related to them.

(c) Degree of nodes, divided by their number of publications, as a function of the diseases they have been related to.

Fig. 2: Correlation with gene morbidity.

TABLE I: Residuals of the linear model

Min	1Q	Median	3Q	Max
-2.25	-1.43	-0.03	1.43	2.58

The residues appear to be normally distributed, with a p-value of 0.81 in a two-sided Kolmogorov-Smirnov test.

TABLE II: Coefficients of the linear model

	Estimate	Std. Error	t value	Pr($>\|t\|$)
β	-1.75	1.48	-1.18	$2.70e^{-1}$
α_1	$3.73e^{-3}$	$9.60e^{-4}$	3.89	$3.67e^{-3}$
α_2	$8.29e^{-2}$	$3.41e^{-2}$	2.43	$3.77e^{-2}$

Both the average degree and the average number of publications per disease are statistically significant. However the number of publications has a significance an order of magnitude higher than the degree. The linear regression has a p-value of $1.41e^{-3}$.

of 0.81) indicate that the linear model is applicable to the data. The low p-value of the model ($1.41e^{-3}$) also suggests that the model fits well the data. Fig. 3 shows some quality measures that strengthen the confidence in the results of the linear model. Fig. 3a shows that the standardized residuals fit the theoretical quantiles relatively well. In addition, Fig. 3b shows that all data points have a low Cook distance, meaning that none of them is causing an important bias in the slope of the regression line [13].

The statistical significance is one order of magnitude higher for the number of publications than for the degree of genes. This suggests that the effect produced by the variance in the amount of information available for the genes is more significant than the effect produced by the variance in the degrees of the genes. Interestingly, the p-value for the degree is yet significant. This means that when the variance in the amount of information is controlled there is still a considerable effect explained in the response variable. In addition, the coefficient calculated for the degree variable is an order of magnitude higher, meaning that given the same amount of information on two genes, the number of diseases they are related with grows at a relative high rate with their degree.

IV. DISCUSSION

Our results show that when the number of publications is not taken into account, there seems to exist a clear relationship between node degree and gene morbidity. However, when the number of publications is in the model, this relationship is not as evident.

There might be an inherent bias in PPI data due to the variance in the amount of information available for different genes. Genes related to diseases appear more in the literature, since they are more interesting to clinical science. In addition, their interacting partners are more thoroughly discovered, since they are the most evident targets for searching disease modulators or new candidate genes. Even though the results suggest a positive correlation between the degree of a node and its morbidity we believe this correlation has to be taken with caution for it can be affected by other factors not taken into account in this study.

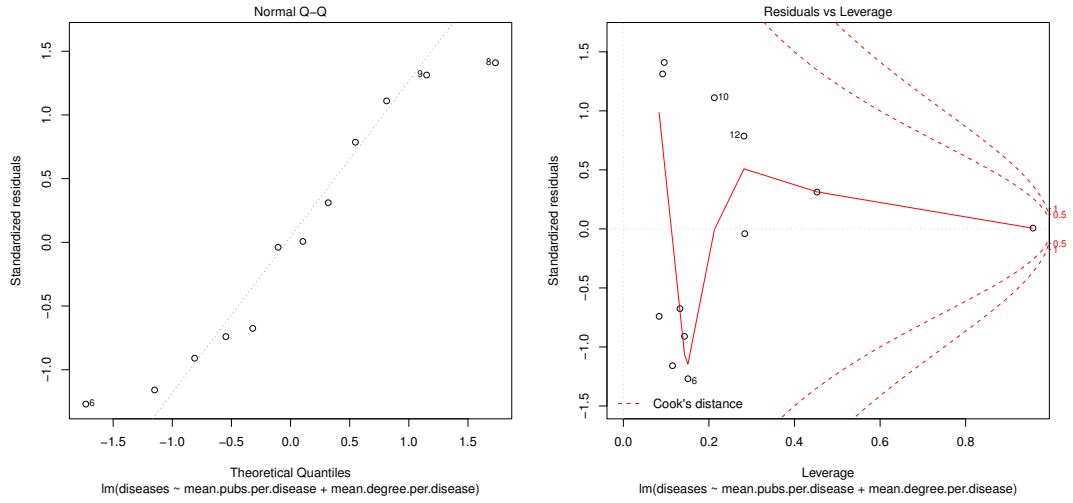
V. ACKNOWLEDGMENTS

The authors want to acknowledge the support received from the Spanish Ministerio de Educación y Ciencia under the Ramón y Cajal Program and TEC2007-63637/TCM and by the Instituto de Salud Carlos III under the initiative CIBER-BBN in Bioengineering, Biomaterials and Nanomedicine.

The authors are very tankful to the reviewers for their many enriching comments.

REFERENCES

[1] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg, "A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae," *Nature*, vol. 403, no. 6770, pp. 623–627, 02/10 2000.



(a) (b)

Fig. 3: (a) The residuals of the model as a function of the predicted values. (b) Cook distance plot of the model.

[2] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S.-L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sorensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. V. Hogue, D. Figeys, and M. Tyers, "Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry," *Nature*, vol. 415, no. 6868, pp. 180–183, 01/10 2002.

[3] A.-C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A.-M. Michon, C.-M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.-A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga, "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature*, vol. 415, no. 6868, pp. 141–147, 01/10 2002.

[4] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal, "Towards a proteome-scale map of the human protein-protein interaction network," *Nature*, vol. 437, no. 7062, pp. 1173–1178, 10/20 2005.

[5] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrín-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. Wu, C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. J. Rilstone, K. Gandi, N. J. Thompson, G. Musso, P. S. Onge, S. Ghanny, M. H. Y. Lam, G. Butland, A. M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili, and J. F. Greenblatt, "Global landscape of protein complexes in the yeast saccharomyces cerevisiae," *Nature*, vol. 440, no. 7084, pp. 637–643, 03/30 2006.

[6] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, 05/03 2001.

[7] N. Przulj, D. A. Wigle, and I. Jurisica, "Functional topology in a network of protein interactions," *Bioinformatics*, vol. 20, no. 3, pp. 340–348, February 12 2004.

[8] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, "The large-scale organization of metabolic networks," *Nature*, vol. 407, no. 6804, pp. 651–654, 10/05 2000.

[9] V. A. McKusick, "Mendelian inheritance in man and its online version, omim," *American Journal of Human Genetics*, vol. 80, no. 4, pp. 588–604, Apr 2007.

[10] T. S. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. H. Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey, "Human protein reference database–2009 update," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D767–772, January 1 2009.

[11] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, Mar. 1947.

[12] R Development Core Team, "R: A language and environment for statistical computing," 2009. [Online]. Available: http://www.R-project.org

[13] R. D. Cook and S. Weisberg, *Residuals and influence in regression*. New York: Chapman and Hall, 1982.